Report from Dagstuhl Seminar 23372

Human-Centered Approaches for Provenance in Automated Data Science

Anamaria Crisan*[1], Lars Kotthoff*[2], Marc Streit*[3], and Kai Xu*[4]

1 **Tableau Software – Seattle, US**
2 **University of Wyoming – Laramie, US**
3 **Johannes Kepler Universität Linz, AT**
4 **University of Nottingham, GB**

Abstract

The scope of automated machine learning (AutoML) technology has extended beyond its initial boundaries of model selection and hyperparameter tuning and towards end-to-end development and refinement of data science pipelines. These advances, both theoretical and realized, make the tools of data science more readily available to domain experts that rely on low- or no-code tooling options to analyze and make sense of their data. To ensure that automated data science technologies are applied both effectively and responsibly, it becomes increasingly urgent to carefully audit the decisions made both automatically and with guidance from humans.

This Dagstuhl Seminar examines human-centered approaches for provenance in automated data science. While prior research concerning provenance and machine learning exists, it does not address the expanded scope of automated approaches and the consequences of applying such techniques at scale to the population of domain experts. In addition, most of the previous works focus on the automated part of this process, leaving a gap on the support for the sensemaking tasks users need to perform, such as selecting the datasets and candidate models and identifying potential causes for poor performance.

The seminar brought together experts from across provenance, information visualization, visual analytics, machine learning, and human-computer interaction to articulate the user challenges posed by AutoML and automated data science, discuss the current state of the art, and propose directions for new research. More specifically, this seminar:

- articulates the state of the art in AutoML and automated data science for supporting the provenance of decision making,
- describes the challenges that data scientists and domain experts face when interfacing with automated approaches to make sense of an automated decision,
- examines the interface between data-centric, model-centric, and user-centric models of provenance and how they interact with automated techniques, and
- encourages exploration of human-centered approaches; for example leveraging visualization.

Seminar Spetember 10–15, 2023 – https://www.dagstuhl.de/23372

2012 ACM Subject Classification Human-centered computing → Visualization; Information systems → Data provenance; Theory of computation → Mathematical optimization; Human-centered computing → Human computer interaction (HCI); Computing methodologies → Machine learning; Computing methodologies → Search methodologies

Keywords and phrases Dagstuhl Seminar, Provenance, AutoML, Data Science, Information Visualisation, Visual Analytics, Machine Learning, Human-Computer Interaction

Digital Object Identifier 10.4230/DagRep.13.9.116

* Editor / Organizer

Human-Centered Approaches for Provenance in Automated Data Science, *Dagstuhl Reports*, Vol. 13, Issue 9, pp. 116–136

Editors: Anamaria Crisan, Lars Kotthoff, Marc Streit, and Kai Xu

Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Anamaria Crisan (Tableau Software – Seattle, US)
Lars Kotthoff (University of Wyoming – Laramie, US)
Marc Streit (Johannes Kepler Universität Linz, AT)
Kai Xu (University of Nottingham, GB)

This Dagstuhl Seminar brings together an interdisciplinary group of researchers and practitioners, spanning Data Science (DS) and Machine Learning (ML), Visualization and Human-Computer Interactions (HCI), and Provenance; to tackle the challenges in automated data science (AutoDS). We specifically focused on ways that methods from human-centered design approaches and provenance can be leveraged to "open up the black box" of AutoDS, introduce greater observability of these methods, and promote human-machine teaming. We observed that there exist many parallel efforts across different disciplines that have yet to be integrated; our seminar brought together these different perspectives as a first step towards producing a general synthesis of methodologies and techniques for advancing AutoDS.

Primitives for AutoDS and hybrid modes of automation. Initial implementations of AutoDS tooling were focused on the so-called CASH problem, combining algorithm selection with parameter optimization, which was exclusively limited to the modeling phase of the data science workflow. More recent work has expanded the scope to include tasks pertaining to data preparation, feature engineering, even model deployment and monitoring for concept drift. Within this expanded end-to-end scope for AutoDS, the individual components of the data science pipeline are often referred to as data science primitives; whether those primitives concern work carried out by a human (i.e., selecting a data set for analysis) or a machine (i.e., hyperparameter tuning) depends on the implementation of the system. Discussions on these data science primitives and the scope of the hybrid automation, where humans and automated processes trade-off work, help frame a discussion around provenance and human-centered design.

Provenance modalities in an end-to-end AutoDS pipeline. Existing methodologies for provenance in data analysis focus on three related themes: data provenance, computation provenance, and user provenance. These are often studied separately, while they should be explored together in AutoDS to be fully transparent and auditable. It was identified that modalities of capturing data, computation, and user provenance may not always align and there exist few techniques that attempt their integration. Moreover, user provenance can be especially complex to capture and surface, as the thinking and reasoning behind analysis choices and decisions are much more challenging to capture than data science workflow or user interactions. Many open problems and potential solutions were discussed at the seminar and more details are provided in the following sections.

Visual and interaction techniques for explainable AutoDS (i.e., model-to-human communication). Data visualization is a powerful medium to help users understand and analyze complex data (in our case the AutoDS provenance), as well as to create opportunities for domain experts and data scientists to interrogate the pipelines themselves. Visual techniques for provenance of AutoDS pipelines exist (i.e., PipelineProfiler, ATMSeer, ModelLineUpper, AutoVizAI, and Visus) but these focus almost exclusively on modeling and do not consider the broader scope of AutoDS primitives. Seminar participants explored the possibilities and utility of visualizing multiple provenance modalities and across AutoDS primitives to achieve this goal.

Human-centered approaches to data science and analytics (i.e., human-to-model communication). Seminar participants acknowledged that humans and automated processes must collaborate in AutoDS, and it becomes necessary to explicitly consider the needs of humans to understand and intervene. Human-centered design encapsulates a broad set of methodologies and techniques for designing technology that interfaces with people. Seminar participants advocated for a broader application in human-centered approaches to ML/AI, including mitigating concerns of "black box" algorithms as discussed earlier. A related research challenge identified is to make DS models more "interactive" so user expertise and knowledge can be more easily incorporated, especially for non-technical domain experts. This can happen during the training of a large model through user "steering" to reduce training time, or after deployment with techniques such as "active learning" to continuously improve the module.

2 Table of Contents

3 Overview of Talks

3.1 Overview of Provenance and Visualization

Kai Xu (University of Nottingham, GB) and Marc Streit (Johannes Kepler University Linz, AT)

In this 30-minute overview talk, we aim to provide a summary of the state-of-the-art of the research related to provenance and its application in interactive visualization. We started with an introduction of what provenance is and how the concept is used with data analysis and visualization. When going through the latest research. We group the work by the "why" (the goal of provenance analysis), the "what" (what provenance data is needed for the intended goal), and "how", (how to capture and analyze the captured provenance). We conclude the talk with a list of open challenges that are important to the field and need further investigation.

3.2 An Introduction to AutoML

Lars Kotthof (University of Wyoming – Laramie, US)

Automated machine learning makes state-of-the-art machine learning accessible to people with little to no background in it. Even for machine learning experts, automated methods are helpful to achieve the best performance with relatively little human effort. In this talk, I will give a high-level overview of the problems that automated machine learning solves and how, after a formal definition of the AutoML problem, I will sketch current solution approaches, issues, open challenges, and potential for application of visualization and provenance approaches.

3.3 Automating Data Science: Pipe Dream or Reality?

Anamaria Crisan (Tableau Research – Seattle, US)

The lack of data scientists but the desire to analyze large data repositories has spurred the development of methods to automate data science. However, in practice, it is complex to orchestrate related human, model, and data processes. Moreover, it becomes difficult to understand how a decision was made and whether this was done by a human or automated process. In this talk, I provide an overview of research motivating the needs and uses of automation. I discuss the existing techniques and tools as well as their limitations. Finally, I discuss the potential of new types of models (specifically LLMs) to further the automation of data science.

3.4 Co-Adaptive Analytics and Guidance

Mennatallah El-Assady (ETH Zürich, CH)

Mixed-initiative visual data analysis systems rely on a process of co-adaptation where human and AI agents collaboratively perform data-driven problem-solving and decision-making. The co-adaptive process describes the dynamic learning and teaching process these agents are engaged in during their interaction in the mixed-initiative system. In this talk, I give an overview of the state-of-the-art in co-adaptive analysis, highlighting co-adaptive guidance in visual analytics. Structuring the topic further, I present the recent paper on deriving a guidance typology. To illustrate how such theoretical concepts can be put into practice, I present two interactive approaches for topic model refinement that employ different types of guidance: speculative execution and single-objective agents. Furthermore, I demonstrate the Lotse library as a practical framework for co-adaptive guidance implementation. Lastly, I discuss open questions concerning provenance, AutoML, and evaluation.

3.5 Exploring Relationships Between Vis/HCI Theory & Provenance

Leilani Battle (University of Washington – Seattle, US)

Visualization theory is often developed at a high level, such as in the form of diagrams, taxonomies, or flow charts. However, these artifacts are difficult to implement in visualization tools. By applying taxonomies to provenance data, such as interaction logs, we could better understand how to make visualization theory more practical.

3.6 DeepCAVE: A visualization and Analysis Tool for AutoML

Tanja Tornede (Leibniz University Hannover, DE)

Joint work of Tanja Torned, René Sass, Eddie Bergman, André Biedenkapp, Frank Hutter, Marius Lindauer
Main reference René Sass, Eddie Bergman, André Biedenkapp, Frank Hutter, Marius Lindauer: "DeepCAVE: An Interactive Analysis Tool for Automated Machine Learning", arXiv, 2022.
URL https://doi.org//10.48550/ARXIV.2206.03493

Visualizing the process of AutoML and its analysis can be done using DeepCAVE. Besides providing a summary of the experimental setup, it offers methods for objective analysis, budget analysis (in multi-fidelity settings), and hyperparameter analysis, This way, the entire interactive framework allows to efficiently generate insights for AutoML problems and brings the human back in the loop.

3.7 Provenance Embedding

Kai Xu (University of Nottingham, GB), Marc Streit (Johannes Kepler Universität Linz, AT)

Joint work of Conny Walchshofer, Andreas Hinterreiter, Kai Xu, Holger Stitz, Marc Streit
Main reference Conny Walchshofer, Andreas P. Hinterreiter, Kai Xu, Holger Stitz, Marc Streit: "Provectories: Embedding-Based Analysis of Interaction Provenance Data", IEEE Trans. Vis. Comput. Graph., Vol. 29(12), pp. 4816–4831, 2023.
URL https://doi.org//10.1109/TVCG.2021.3135697

In this talk, we propose a research question that may be of interest to seminar participants for discussion during the seminar. The idea is based on a previous work on modeling and visualizing provenance. The main idea is to capture provenance as a vector sequence, which can then be visualized and analyzed using techniques designed for high-dimensional data. The new idea is to take this one step further, following the process similar to training Large Language Models (LLM) such as ChatGPT by masking a step in the provenance and training a model to predict it. The hope is that such a model can have additional 'intelligence' besides predicting the next step, similar to ChatGPT.

3.8 Trrack + Persist

Kiran Gadhave (University of Utah – Salt Lake City, US)

Joint work of Zach Cutler, Kiran Gadhave, Alexander Lex
Main reference Zach Cutler, Kiran Gadhave, Alexander Lex: "Trrack: A Library for Provenance-Tracking in Web-Based Visualizations", in Proc. of the 31st IEEE Visualization Conference, IEEE VIS 2020 – Short Papers, Virtual Event, USA, October 25-30, 2020, pp. 116–120, IEEE, 2020.
URL https://doi.org//10.1109/VIS47514.2020.00030

Trrack is a provenance tracking library for the web. One of the goals of the library is to be easy to integrate Trrack. Trrack has a hybrid provenance tracking approach which tracks both actions and state. Trrack stores the diffs between the states to optimize storage.

Computational notebooks have a gap between code and visualization. Semantic, layout and temporal gap are the three highlighted in B2 by Wu et al. [1]. B2 proposes queries (e.g. elections) as a bridge between them. We propose using Trrack provenance to bridge the gap. We've done a Jupyter extension which shows examples of this.

References

1 Yifan Wu, Joseph M. Hellerstein, and Arvind Satyanarayan. B2: Bridging Code and Interactive Visualization in Computational Notebooks. *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, 152–165, 2020, doi: 10.1145/3379337.3415851.

3.9 Mosaic

Dominik Moritz (Carnegie Mellon University – Pittsburgh, US)

Joint work of Jeffrey Heer, Dominik Moritz
Main reference Jeffrey Heer, Dominik Moritz: "Mosaic: An Architecture for Scalable & Interoperable Data Views", IEEE Trans. Vis. Comput. Graph., Vol. 30(1), pp. 436–446, 2024.
URL https://doi.org//10.1109/TVCG.2023.3327189

Mosaic is an extensible framework for linking databases and interactive views. It links charts, tables, inputs, etc. through a coordinator that optimizes queries and creates data cube indices (for fast linked interactions) as tables in the database. Mosaic is very useful for building linked dashboards and in the future we could track provenance to log it in studies or suggest analyses or subspaces of the data to look at.

3.10 Understanding How In-Visualization Provenance Can Support Trade-off Analysis

Mehdi Chakhchoukh (Université Paris-Saclay, FR)

Joint work of Mehdi Chakhchoukh, Nadia Boukhelifa, Anastasia Bezerianos
Main reference Mehdi Chakhchoukh, Nadia Boukhelifa, Anastasia Bezerianos: "Understanding How In-Visualization Provenance Can Support Trade-Off Analysis", IEEE Trans. Vis. Comput. Graph., Vol. 29(9), pp. 3758–3774, 2023.
URL https://doi.org//10.1109/TVCG.2022.3171074

In domains such as agronomy or manufacturing, experts need to consider trade-offs when making decisions that involve several, often competing, objectives. Such analysis is complex and may be conducted over long periods of time, making it hard to revisit. In this talk we presented some of our results that were published in an IEEE Transactions on Visualisation and Computer Graphics paper: mainly the idea of refining Ragan et al. [1] purposes for provenance with provenance objects that are task-specific. We discussed if such objects could be used to support the design of provenance visualization for autoML tasks. Finally, we presented the challenges encountered when designing provenance views based on our experience from the experiments we ran with agronomy experts with real-world data and applications.

References

1 Eric D. Ragan, Alex Endert, Jibonananda Sanyal and Jian Chen. Characterizing Provenance in Visualization and Data Analysis: An Organizational Framework of Provenance Types and Purposes. *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 31-40, 31 Jan. 2016, doi: 10.1109/TVCG.2015.2467551.

3.11 Data Provenance for Reproducible Research

Sheeba Samuel (Friedrich Schiller University – Jena, DE)

Joint work of Sheeba Samuel, Daniel Mietchen
Main reference Sheeba Samuel, Daniel Mietchen: "Computational reproducibility of Jupyter notebooks from biomedical publications", CoRR, Vol. abs/2209.04308, 2022.
URL https://doi.org//10.48550/ARXIV.2209.04308

Reproducible research refers to the idea that scientific results are documented and published in a way that others may verify the findings and build upon them. Data provenance is one of the integral components of reproducible research across domains. We investigate the computational reproducibility aspect of Jupyter notebooks within the context of publications indexed in PubMed Central. Our research endeavors to identify common challenges and best practices, delineate emerging trends and propose potential enhancements to Jupyter-related workflows associated with publications. To bolster the reproducibility of Jupyter-related workflows, we delve into various data provenance approaches and tools. Specifically, we examine the utility of tools such as ProvBook [1] and MLProvLab [2] in capturing and visualizing diverse aspects of provenance information. MLProvLab, in particular, enables granular tracking of information at both the notebook and cell levels, visualizing dependencies between cells and data within a notebook. This functionality is invaluable for data scientists, as it aids in comprehending the cascading effects of changes made to one cell on subsequent cells and, ultimately, on the research results. Furthermore, we revisit the W3C model, PROV-O [3], for representing provenance information, emphasizing the pivotal role of ontologies in modeling such information effectively. Our exploration extends to the ReproduceMe data model, which facilitates the sharing of computational provenance in a machine-readable format, enhancing the accessibility and utility of provenance information. Finally, we address research questions concerning the significance of provenance information and its utilization in machine learning and deep learning pipelines. We underscore the importance of sharing and harnessing collected provenance information to enhance the transparency, reproducibility, and trustworthiness of research outcomes in these domains.

References

1 Sheeba Samuel and Birgitta König-Ries. ProvBook: Provenance-based Semantic Enrichment of Interactive Notebooks for Reproducibility, 17th International Semantic Web Conference (ISWC) 2018 Demo Track.
2 Dominik Kerzel, Sheeba Samuel, and Birgitta König-Ries. Towards Tracking Provenance from Machine Learning Notebooks. *Proceedings of the 13th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management* , 274-281, 2021, doi: 10.5220/0010681400003064.
3 Paolo Missier, Khalid Belhajjame, and James Cheney. The W3C PROV family of specifications for modelling provenance metadata. *Proceedings of the 16th International Conference on Extending Database Technology*, pages 773–776, New York, USA, 2013.

3.12 Welcome to Parameter Land – Visual Parameter Space Exploration

Klaus Eckelt (Johannes Kepler University Linz, AT)

URL https://observablehq.com/@keckelt/dekumap

We introduce an approach to visualize and navigate the hyperparameter space of machine learning pipelines optimized by AutoML methods. Rather than parallel coordinates plots, we envision the hyperparameter space as a dynamic map, allowing users to intuitively explore, optimize, and discover regions of interest–or simply monitor the optimization process. Providing interactive visualizations, like treemaps or LineUp [1], should ultimately enhance the user experience in AutoML leading to higher trust and informed decision-making in machine learning pipelines.

References

1 Samuel Gratzl, Alexander Lex, Nils Gehlenborg, Hanspeter Pfister, and Marc Streit. LineUp: Visual Analysis of Multi-Attribute Rankings. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2277–2286, 2013, doi: 10.1109/TVCG.2013.173.

4 Working Groups

4.1 Terminology

Alex Endert (Georgia Tech – Atlanta, US) Alexander Lex (Utah University – Salt Lake City, US) Alvitta Ottley (Washington University in St Louis, US) Ana Crisan (Tableau Research, Salesforce – Seattle, US) Camelia D. Brumar (Tufts University – Medford, US) Kai Xu (University of Nottingham, GB) Leilani Battle (University of Washington – Seattle, US) Marc Streit (Johannes Kepler University Linz, AT) Menna El-Assady (ETH Zürich, CH) Nadia Boukhelifa (Université Paris-Saclay, FR)

Embarking on a journey through the multifaceted domain of data provenance, this working group unfolded discussions and explorations, intertwining the conceptual, terminological, and applicative aspects of provenance data. The exploration, situated within the realms of Human-AI interaction and adaptive systems, traversed through a meticulous terminological exploration and an analytical discourse, rooted in a transcript reflecting upon the dual roles of providence data in adaptive systems. The commitment was not just to comprehend but to unfold and apply this understanding in real-world scenarios, especially where the human and artificial intelligence amalgamate.

The terminological exploration unwrapped the concept of "provenance data", probing its constituents, potential applications, and ontological management, with a nuanced focus on existing frameworks like Google's ontologies. The discussions, while illuminating, also underscored the imperative for a cohesive and adaptable understanding of provenance data, which can seamlessly weave through varied applications and domains, ensuring its utility across a spectrum of use-cases, especially within the intricate fabric of Human-AI interactions. A parallel strand navigated through the active and passive roles of providence data within

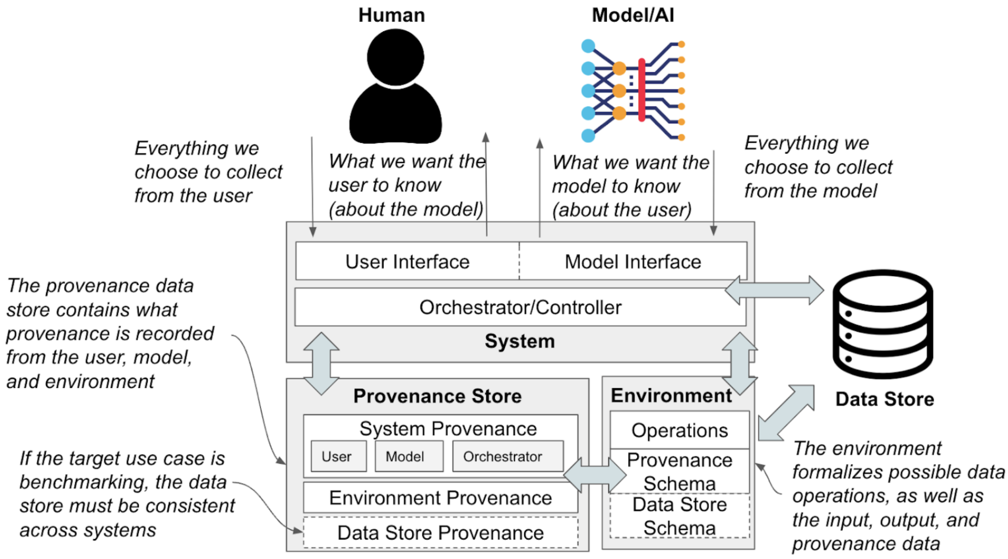

Figure 1 A conceptual model illustrating the interplay between human actors, artificial intelligence, and a visual analytics system designed to capture provenance data.

adaptive systems and machine learning, spotlighting its potential to not only train models through human interactions but also to elucidate model behaviors, crafting a bi-directional pathway of understanding and application.

Emerging from the discourse was the ontological challenge and a palpable paradox: the need to comprehensively comprehend and define provenance data while concurrently delving into its applications and management in pragmatic contexts. This paradox, particularly pronounced in discussions around adaptive systems, spotlighted the necessity for a clear, structured understanding as pivotal to harnessing providence data's full potential. Simultaneously, it brought forth the challenges and gaps extant in leveraging ontological frameworks across diverse applications and domains, necessitating further exploration and refinement to make these frameworks universally adaptable and coherent.

As the group forges ahead, the commitment is twofold: refining and expanding the understanding and applications of provenance data and ensuring that this theoretical clarity is translatable into pragmatic applications, especially in crafting intuitive, transparent, and effective adaptive systems. This involves not only a deeper exploration and defining of the terminology and conceptual frameworks but also a meticulous examination of its applications, ensuring a seamless transition from theory to practice. Additionally, a continuous, collaborative dialogue with the wider academic and research community is envisioned, wherein the group not only shares its findings and insights but also invites perspectives, critiques, and contributions, ensuring a holistic, multifaceted approach towards understanding and harnessing provenance data effectively.

This section attempts to weave the discussions, explorations, and future directions into a coherent narrative, based on the initial understanding from the provided text files. If there are specific aspects or nuances you'd like to explore or emphasize further, please provide additional guidance or specify areas of deeper interest.

4.1.1 Discussions on Definition

This part of the discussion focused on the questions "What really is provenance data?" and "how does it relate to similar concepts?". This is broken down into a few sub-questions:

1. What can be considered provenance?
2. What are the major use cases to think about for provenance data?
3. Is it related to ontologies? Google has its ontologies, which is a valuable asset for improving product recommendations. Maybe this is also related to the knowledge graphs or knowledge bases?
4. Could we integrate knowledge graphs for visualization recommendation/visual analysis?
5. Maybe this is also related to "Grammars" that formalizes how researchers process and reason about provenance data
   a. Reconciling coarser and finer levels of abstraction for provenance data
   b. Could we formulate grammars to represent interactions?

In the context of human-AI teaming, there are three main provenance components. All of these need to be captured to provide a complete provenance.

1. User's reasoning/mental model (including granularity)
2. ML's reasoning/"mental" model (including granularity)
3. Communication between human and ML agents

The availability of provenance has a large impact on possible downstream tasks. The "Imperative" provenance, such as sequences of user interactions, is usually easy to capture, but has limited semantics. "Declarative" provenance, such as user goals, is often more difficult to capture thus less available, but they provide useful insight into the analysis process. The group also observed that there are differences in definitions between the VIS and ML community, and this is to some extent decided by the capabilities of the tool at hand, e.g., limiting supported tasks, available interactions, etc.

4.1.2 A possible "Opinionated" Survey Paper

The group had a long discussion about writing an overview or survey paper and decided to start with some possible sections the paper would have and what will be covered in each section. These are detailed below, together with other aspects of the paper the group considered.

4.1.2.1 Preliminaries

These are the assumptions the all the discussions will be based on:

- We consider three main components: humans (users), AI/models, and system ("environment")
- The provenance captures the history of the "environment":
  1. Where the human performs the actions the model cares about;
  2. In our case, this is likely an analysis UI/system;
  3. Environment vs Interface vs System: this is a core concept of the paper and the group spend a long time discussing it, which led to the conceptual model shown in Figure 1.

4.1.2.2 Paper search methods

The group then discussed various publication collections that the survey will cover:
1. Review proceedings of VIS, EuroVis, CHI, IUI, FAccT, CG&A, TVCG, TiiS, IVS, CGF,
2. Google scholar search (make sure we record the exact keyword searches we used)
3. LLM search: Abstracts Viewer (https://pages.graphics.cs.wisc.edu/AbstractsViewer/) and http://vitality.mathcs.emory.edu/app
4. Papers (metadata) collected by Shayan (the conferences and journals covered): AAAI, AISTATS, CGF, CHI, ICML, IUI, KDD, NeurIPS, TiiS, TVCG, VIS (https://github.com/smonadjemi/mixed-initiative-va-survey/tree/main/paper_data)
5. Publications from provenance conferences such as Theory and Practice of Provenance

4.1.2.3 What do we consider as provenance?

This closely relates to the scope of the paper. While this paper might not cover every type of provenance, it is useful to have a relatively complete list and then decide what to include. "You might not think it's provenance but it actually is." – Ana. It became clear to some of the group members that their work is related to provenance, but they did not realise that because it was described with a different term. For example, user interaction log is a common type of provenance and ubiquitous in studies evaluating different visual designs and visual analysis systems. However, not everyone realises that this is a form of provenance, and this can be commonplace within the community.

Provenance recording is typically done for a purpose. This has been covered in previous work [20, 27] , but there is a need for further discussion. One fundamental issues is the relationships among the subsets of provenance, as shown in Figure 2. There is the "provenance" in the broadest sense and includes everything that can be captures theoretically. Within this, there is provenance that can be captured practically (the "capturable provenance"), and the ones that no effective recording means exists (such as capturing user thinking). Among the "capturable provenance", some of them are "interpretable", i.e., human can make sense of it. Finally, there is the "relevant provenance", which depends on the application and analysis question, that overlaps all the other types. The group also observed that there is difference between the common interests for academia and industry: while the visualization research community is often more interested in interaction and evaluation, industry tend to care more about data quality, model performance, and governance issues.

There are pros and cons to recording different forms of provenance: Passive/automated log recording does not disrupt the user but can be noisy and lack of meaning, whereas Explicit feedback can be higher quality data for models but disrupts the user's flow. "Passive" Interaction logs include raw system event data and user clicking on typical UI components. However, this also consists of implicit feedback, such as selecting among recommendations, which provides information with richer semantics. This relates to the design idea of "dual purpose interactions", i.e. interactions for performing operations and learning about users. Explicit Feedback, i.e., things you directly ask the user, is less common but it can be very helpful to improve the performance of machine learning models [16]. It is also possible to divide the forms of provenance by its source, i.e., provenance about user and provenance about the system/model. Currently there is no system/tool that combines different types of provenance, especially model+ interaction provenance.

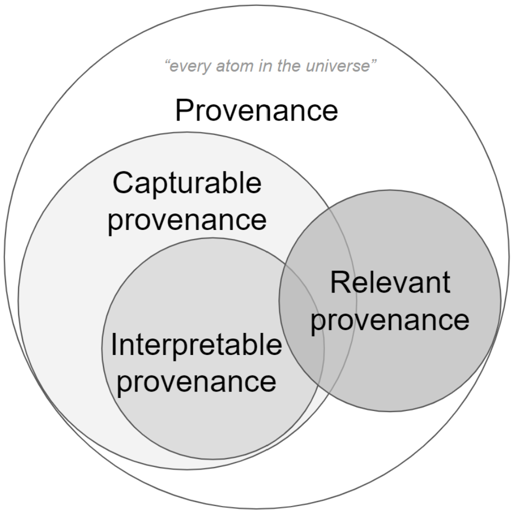

Figure 2 Different types of provenance and their relationships.

4.1.2.4 Human-AI Interaction

Given the topic of this seminar, human-AI interaction is of particular interest to all the participants. This can be broken down further, as shown in Table 1.

Table 1 H = high importance, L = low importance.

Objective	Provenance	ML	User	Data
ML Guidance	XAI, Open model	H	L	
Guiding ML	Learn form user interaction	L	H	
Auditing	Overview, observability	H	H	H
Benchmarking	model performance	H	L	H

4.1.2.5 Representations of Human-AI Interaction

A natural follow-on question is how to represent the provenance of human-AI interaction. Borrowing from the Linguistics community, there can be *semantic* versus *syntactic* relationships between elements. Usually, *grammar* defines syntactic relationships, while *ontology* defines semantic relationships. This has been attempted before for provenance tasks [2], which theorizes a mapping between low-level user interactions and high-level user insights, taking a hierarchical, grammar-like approach. There is also work from outside the visualization community, such as the "Structural summaries for visual provenance analysis" [11] and "Automated Provenance Analytics: A Regular Grammar Based Approach with Applications in Security" [14].

4.1.2.6 Ethical Considerations

There are many ethical considerations related to the collection and use of provenance. One example is *profiling*: the provenance information can be potentially used to profile its source, often a human in this context, for purpose beyond what is originally intended. Related to this is anonymization, i.e., how to remove the personal identifiable information from provenance. An alternative is to seek explicit consent, i.e., user agrees share the information through *data donations*.

4.2 Humans

Jen Rogers (Tufts University – Medford, US) Emily Wall (Emory University – Altanta, US) Mehdi Chakhchoukh (Universite Paris-Saclay, FR) Marie Anastacio (Leiden Universiteit, NL) Rebecca Faust (Virginia Tech – Blacksburg, US) Cagatay Turkay (University of Warwick, GB) Lars Kotthoff (University of Wyoming – Laramie, US) Steffen Koch (University of Stuttgart, DE) Andreas Kerren (Linköping University, Sweden) Jürgen Bernard (University of Zürich, CH)

This working group had a total of 10 attendees, who brought diversified perspectives, enriching the discourse. Among the attendees were experts in autoML, and experts in data visualisation. The group focused its efforts on the trade-offs between humans and automation within automated data science. Key areas explored included understanding the AutoML "black box", the role of provenance in supporting diverse user cases, the challenges and affordances of humans and automation, the significance of visualization, and trust-building within AutoML.

The autoML experts showed particular interest in the human interaction aspect at the meta-level, sparking a debate and a categorisation of the main failures that could arise while utilizing autoML. During this categorisation we identified the failures that stem from the automated part of autoML as well as the failures that could stem from the human beings. For instance, is a model's subpar performance an inherent problem to the autoML process or a setup error that comes from human's input?

This led to a deeper discussion about how to unveil the black box surrounding the model and where to include the human in the autoML pipeline. These discussions highlighted provenance as vital to support different user needs, including refinement of user tasks and capturing the rationale for chosen models and their subsequent outcomes. However, further reflection determined that to speculate on provenance, we first needed to define key roles, affordances, challenges of humans, and automation within the data science pipeline as we viewed this as a necessary foundation for further speculation.

Visualization is crucial for helping users understand the process, decisions, and trade-offs within the data science pipeline. Despite its importance in comprehension and trust, the group acknowledged the challenges in visualizing aspects of the process. Additionally, the role of visualization was noted to differ among users, necessitating a nuanced approach tailored to specific stakeholders.

The paradox of human involvement in automated data science remained a consistent theme during our discussions (Figure 3). Do we need humans in the loop at all? What value do they contribute to the data science pipeline? Can we trust a process that we do not understand and are not involved in? Questions such as these provided a catalyst for the group's current efforts in defining tradeoffs between human and automation along the data science pipeline.

Moving forward, the group will continue to develop and refine its perspective on this paradox. This development will include collecting data from the community on their opinions of current and future human-to-machine balances within the data science pipeline through an anonymous online survey. The group will formalize its findings into a written report to contribute to the wider academic community. A potential publication outlet would be the IEEE Computer Graphics and Applications (CG&A) journal.

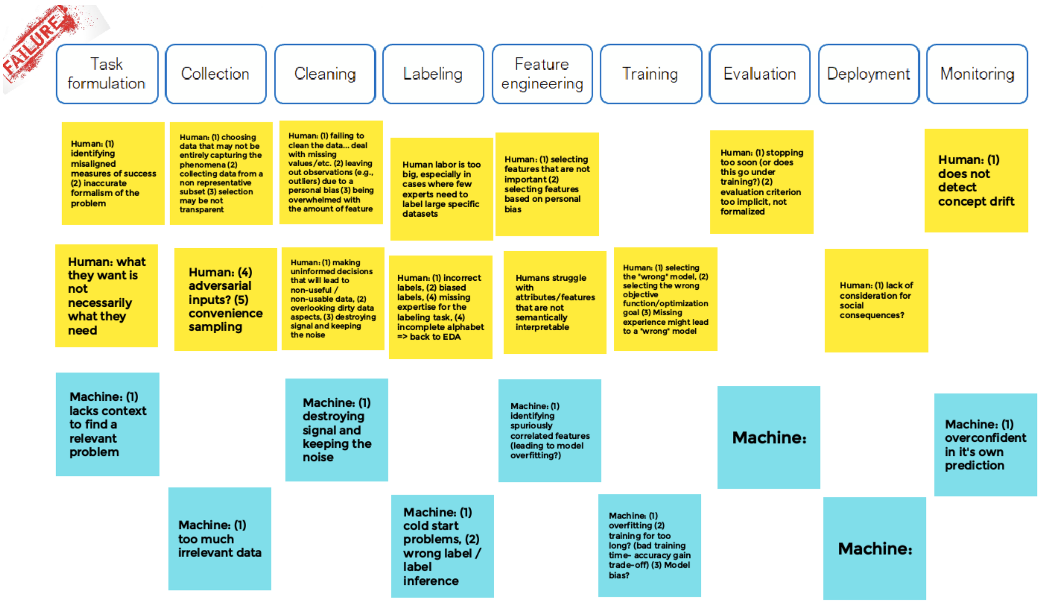

Figure 3 The Human Paradox of AutoML.

4.3 Applications

Klaus Eckelt (Johannes Kepler University Linz, AT), Sheeba Samuel (Friedrich Schiller University Jena, DE), David Koop (Northern Illinois University – DeKalb, US), Kiran Gadhave (Utah University – Salt Lake City, US), Dominik Moritz (Carnegie Mellon University – Pittsburgh, US)

Automated Machine Learning (AutoML) and Data Science (AutoDS) pipelines have gained significant attention in recent years. This working group focused on interactive systems to track and visualize the provenance of these pipelines and enable user interaction with automated algorithms as they are running. To gain an overview of the state-of-the-art, we reviewed related work that compares AutoML/DS libraries [5] and applied multiple of them to the same tabular data set. We analyzed the information they provide on the ongoing optimization process, the search space, and their final result (see Table 2). We also compared the resulting models by their complexity and accuracy. All the analyzed libraries provide logs in the console and optionally in a file. Our analysis included Auto-sklearn [7, 6], AutoGluon [4], TPOT [12], FLAML [25], and H2O AutoML [13]. These libraries were used within Jupyter Notebook using Python.[1]

Although some AutoML/DS users have extensive domain knowledge, they may have limited knowledge in the realm of machine learning, and conversely, individuals with a strong machine learning background may lack expertise in the specific domain of the data [3]. As a result, different user groups require different levels of detail and information in visualizing the AutoML/DS process. Related work also differs in terms of detail presented and potential target users. While partial dependence or parallel coordinate plots are easily interpretable,

[1] https://github.com/keckelt/dagstuhl-23372-applications

Table 2 Overview of the considered AutoML libraries. Downloads were retrieved from PyPI Stats for the last 30 days [8]. The runtime limit was set to 30 minutes. All pipelines were trained with default settings, but maximized CPU utiliznation and logging outputs (in the console (>_) and file (🗎)). ∗ CASH. . . Combined Algorithm Selection and Hyperparameter Optimization; † PoSH. . . Portfolio Successive Halving.

Library	Downloads	Log	Search & Optimization Strategy	Accuracy
Auto-sklearn	25,000	>_ 🗎	CASH* with Bayesian Optimization	95.35 %
Auto-sklearn 2.0	25,000	>_ 🗎	PoSH† and Bayesian Optimization	93.07 %
AutoGluon	41,000	>_ 🗎	Model Portfolio and Random Search	94.30 %
TPOT	35,000	>_ 🗎	CASH with Genetic Programming	99.92 %
FLAML	189,000	>_ 🗎	CASH with Cost-Frugal Optimization	98.89 %
H2O AutoML	340,000	>_	CASH with Random Search	94.44 %

tools such as PipelineProfiler [19], ATMSeer [26], or DeepCave [21] allow for a more detailed inspection, but also require technical understanding of the parts of the machine learning pipeline and its optimization.

AutoML/DS approaches and the tools to visualize them currently provide little room for human interaction. The only way to steer the algorithms is by setting parameters before starting an(other) optimization run. But for optimal performance, AutoGluon, for example, advises against human intervention in its documentation[2] and Auto-sklearn 2.0 also removed the human from the loop [6]. Even more steps of the AutoML/DS pipeline will be automated in the future [10].

However, recent research argues for the necessity of reintegrating users into the loop [15]. Given that AutoML/DS requires time to run, it is essential to allow adaptations to make efficient use of this time. Interactive visualization systems can help identify issues early on. For instance, a label left in the training data could cause unusually high performance across all configurations, or poor performance could be due to insufficient data quality. These systems would also enable users to make adjustments to the performance metric, to trade-off between sensitivity and specificity, for example. Providing users with more control when necessary can increase their trust in the AutoML/DS system, often perceived as a 'black box', and could also speed up and improve the process by allowing users to contribute their knowledge more effectively. Meta-learning – i.e., learning from previous experiments – is currently only supported on the machine side of the loop. Auto-sklearn 2.0 [6], for example, looks for similar problems on OpenML [24] to learn from them. However, AutoML/DS systems do not allow users to provide information on similar problems they have encountered. Allowing users to provide their knowledge to the optimization process could guide the search throughout the optimization process.

Figure 4 shows our sketch to visualize AutoML/DS system processes, compare them, and interact with them. Multiple runs can be selected at the top. Their progress and performance metric is displayed on the left side. The large table on the right gives an overview of the configurations that were trained over time. The individual runs are distinguished through different colors (■, ■, ■,). If a configuration fails, we use the negative of the run's color (■, ■, ■). These negatives are less saturated to better differentiate them from successful runs. As there are more configurations than can be displayed on the available vertical screen space, an aggregation method can be selected using the radio buttons on the top right. The best

[2] `https://auto.gluon.ai/stable/tutorials/tabular/tabular-essentials.html#maximizing-predictive-performance`

Figure 4 Our sketched visual interface to visualize AutoML/DS systems. An ongoing Auto-sklearn optimization is visualized in green. Additional past runs can be selected at the top for comparison. The table on the right shows the best performing configurations in the specific time segment.

overall configuration per run is additionally highlighted with a colored horizontal bar across the entire line and a trophy symbol next to it. For a detailed inspection, the progress bar, accuracy plot, and table can be vertically expanded using a switch button to show each tested configuration without aggregation. This table can also be used for interaction with the AutoML/DS process. Users should be able to prioritize or block elements to be explored in order to guide the search space. Using an interactive table like LineUp [9] would also allow to filter and rank the configurations, and update performance metrics through combination and weighting of recorded information. However, such an interaction is currently not possible in any of the AutoML/DS systems we reviewed.

We also found that none of the AutoML libraries we tried support MLOps services like MLflow [28] or Weights and Biases [1], which are frequently used to track the training and optimization process of machine learning projects. A custom logging configuration can be passed to Auto-sklearn. All other libraries were only able to log into files instead. We wanted to visualize the AutoML/DS process in real-time while it is running, and thus defined our own logger for Auto-sklearn that also sends all output to Weights and Biases.[3] As this approach heavily relies on log data and parsing string outputs it is limited and error prone. We also noted that these MLOps services do not support the process of tracking AutoML/DS optimizations well, due to the pipeline's many different elements and their parameters.

MLflow or Weights and Biases do not store the recorded information in any standardized or interoperable format, such as PROV-ML [23], which is based on W3C PROV [18]. With *mlflow2prov* [22], data from MLflow and the versioned source code from which it originates can be combined into another provenance format based on W3C PROV.

In addition to the AutoML/DS provenance, a common format to describe the tracked data is necessary. Pipeline elements, their naming, and possible combinations vary between AutoML/DS tools. To ensure that interactive systems are interoperable between AutoML/DS

[3] `https://wandb.ai/dagstuhl-23372/automl`

tools, they require a common standard to communicate and store (intermediate) results [17]. This would allow users to visualize and compare the results of different AutoML/DS systems beyond the performance metrics. We argue that AutoML/DS systems require hooks with which intermediate results are communicated and APIs to steer the ongoing process. We believe the adoption of a standardized provenance format, which can be shared between different data science tools, would facilitate comparisons and allow better monitoring, debugging, interpretation, and explanation of the process.

References

1 Lukas Biewald. Experiment Tracking with Weights and Biases. `https://www.wandb.com/`, 2020. Accessed: 2023-09-18.

2 Christian Bors, John Wenskovitch, Michelle Dowling, Simon Attfield, Leilani Battle, Alex Endert, Olga Kulyk, and Robert S. Laramee. A provenance task abstraction framework. *IEEE Computer Graphics and Applications*, 39(6):46–60, 2019.

3 Anamaria Crisan, Brittany Fiore-Gartland, and Melanie Tory. Passing the Data Baton: A Retrospective Analysis on Data Science Work and Workers. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1860–1870, 2021.

4 Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data. *arXiv preprint arXiv:2003.06505*, 2020.

5 Luís Ferreira, André Pilastri, Carlos Manuel Martins, Pedro Miguel Pires, and Paulo Cortez. A Comparison of AutoML Tools for Machine Learning, Deep Learning and XGBoost. In *2021 International Joint Conference on Neural Networks*, pages 1–8, July 2021.

6 Matthias Feurer, Katharina Eggensperger, Stefan Falkner, Marius Lindauer, and Frank Hutter. Auto-sklearn 2.0: hands-free AutoML via meta-learning. *The Journal of Machine Learning Research*, 23(1):261:11936–261:11996, January 2022.

7 Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Springenberg, Manuel Blum, and Frank Hutter. Efficient and Robust Automated Machine Learning. In *Advances in Neural Information Processing Systems 28 (2015)*, pages 2962–2970, 2015.

8 Christopher Flynn. PyPI Download Stats. `https://pypistats.org/`, 2023. Accessed: 2023-09-18.

9 Samuel Gratzl, Alexander Lex, Nils Gehlenborg, Hanspeter Pfister, and Marc Streit. LineUp: Visual Analysis of Multi-Attribute Rankings. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2277–2286, 2013.

10 Shubhra Kanti Karmaker ("Santu"), Md. Mahadi Hassan, Micah J. Smith, Lei Xu, Chengxiang Zhai, and Kalyan Veeramachaneni. AutoML to Date and Beyond: Challenges and Opportunities. *ACM Computing Surveys*, 54(8):175:1–175:36, 2021.

11 Houssem Ben Lahmar and Melanie Herschel. Structural summaries for visual provenance analysis. In *Proceedings of the 11th USENIX Conference on Theory and Practice of Provenance*, TAPP'19, page 2, USA, 2019. USENIX Association.

12 Trang T Le, Weixuan Fu, and Jason H Moore. Scaling tree-based automated machine learning to biomedical big data with a feature set selector. *Bioinformatics*, 36(1):250–256, 2020.

13 Erin LeDell and Sebastien Poirier. H2O AutoML: Scalable Automatic Machine Learning. *7th ICML Workshop on Automated Machine Learning*, July 2020.

14 Mark Lemay, Wajih Ul Hassan, Thomas Moyer, Nabil Schear, and Warren Smith. Automated provenance analytics: A regular grammar based approach with applications in security. In *Proceedings of the 9th USENIX Conference on Theory and Practice of Provenance*, TaPP'17, page 12, USA, 2017. USENIX Association.

15 Marius Lindauer and Alexander Tornede. Rethinking AutoML: Advancing from a Machine-Centered to Human-Centered Paradigm. `https://www.automl.org/rethinking-automl-advancing-from-a-machine-\ centered-to-human-centered-paradigm/`, 2022. Accessed: 2023-09-18.

16 Yannick Metz, David Lindner, Raphaël Baur, Daniel Keim, and Mennatallah El-Assady. RLHF-Blender: A Configurable Interactive Interface for Learning from Diverse Human Feedback, August 2023. arXiv:2308.04332 [cs].

17 Mitar Milutinovic. *Towards Automatic Machine Learning Pipeline Design*. PhD thesis, EECS Department, University of California, Berkeley, Aug 2019.

18 Paolo Missier, Khalid Belhajjame, and James Cheney. The W3C PROV family of specifications for modelling provenance metadata. In *Proceedings of the 16th International Conference on Extending Database Technology*, pages 773–776, New York, USA, 2013. ACM.

19 Jorge Piazentin Ono, Sonia Castelo, Roque Lopez, Enrico Bertini, Juliana Freire, and Claudio Silva. PipelineProfiler: A Visual Analytics Tool for the Exploration of AutoML Pipelines. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):390–400, February 2021.

20 E.D. Ragan, A. Endert, J. Sanyal, and J. Chen. Characterizing Provenance in Visualization and Data Analysis: An Organizational Framework of Provenance Types and Purposes. *IEEE Transactions on Visualization and Computer Graphics (VAST '15)*, 22(1):31–40, 2016.

21 René Sass, Eddie Bergman, André Biedenkapp, Frank Hutter, and Marius Lindauer. DeepCAVE: An Interactive Analysis Tool for Automated Machine Learning. In *ICML Workshop on Adaptive Experimental Design and Active Learning in the Real World*, June 2022.

22 Marius Schlegel and Kai-Uwe Sattler. Extracting Provenance of Machine Learning Experiment Pipeline Artifacts. In Alberto Abelló, Panos Vassiliadis, Oscar Romero, and Robert Wrembel, editors, *Advances in Databases and Information Systems*, Lecture Notes in Computer Science, pages 238–251. Springer Nature Switzerland, 2023.

23 Renan Souza, Leonardo Azevedo, Vítor Lourenço, Elton Soares, Raphael Thiago, Rafael Brandão, Daniel Civitarese, Emilio Brazil, Marcio Moreno, Patrick Valduriez, Marta Mattoso, Renato Cerqueira, and Marco A.S. Netto. Provenance Data in the Machine Learning Lifecycle in Computational Science and Engineering. In *2019 IEEE/ACM Workflows in Support of Large-Scale Science*, pages 1–10, November 2019.

24 Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. OpenML: networked science in machine learning. *SIGKDD Explorations*, 15(2):49–60, 2013.

25 Chi Wang, Qingyun Wu, Markus Weimer, and Erkang Zhu. FLAML: A Fast and Lightweight AutoML Library. *Proceedings of Machine Learning and Systems*, 3:434–447, 2021.

26 Qianwen Wang, Yao Ming, Zhihua Jin, Qiaomu Shen, Dongyu Liu, Micah J. Smith, Kalyan Veeramachaneni, and Huamin Qu. ATMSeer: Increasing Transparency and Controllability in Automated Machine Learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, New York, USA, 2019. ACM.

27 Kai Xu, Alvitta Ottley, Conny Walchshofer, Marc Streit, Remco Chang, and John Wenskovitch. Survey on the Analysis of User Interactions and Visualization Provenance. *Computer Graphics Forum*, 39(3):757–783, 2020. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14035.

28 Matei Zaharia, Andrew Chen, Aaron Davidson, Ali Ghodsi, Sue Ann Hong, Andy Konwinski, Siddharth Murching, Tomas Nykodym, Paul Ogilvie, Mani Parkhe, and others. Accelerating the Machine Learning Lifecycle with MLflow. *IEEE Data Engineering Bulletin*, 41(4):39–45, 2018.

Participants

- Marie Anastacio
RWTH Aachen, DE
- Leilani Battle
University of Washington – Seattle, US
- Jürgen Bernard
Universität Zürich, CH
- Nadia Boukhelifa
INRAE – Palaiseau, FR
- Camelia D. Brumar
Tufts University – Medford, US
- Mehdi Chakhchoukh
Université de Paris-Saclay – Gif-sur-Yvette, FR & INRAE – Paris, FR
- Anamaria Crisan
Tableau Software – Seattle, US
- Klaus Eckelt
Johannes Kepler Universität Linz, AT
- Mennatallah El-Assady
ETH Zürich, CH
- Alex Endert
Georgia Institute of Technology – Atlanta, US
- Rebecca Faust
Virginia Polytechnic Institute – Blacksburg, US
- Kiran Gadhave
University of Utah – Salt Lake City, US
- Andreas Kerren
Linköping University, SE
- Steffen Koch
Universität Stuttgart, DE
- David Koop
Northern Illinois University – DeKalb, US
- Lars Kotthoff
University of Wyoming – Laramie, US
- Alexander Lex
University of Utah – Salt Lake City, US
- Dominik Moritz
Carnegie Mellon University – Pittsburgh, US
- Alvitta Ottley
Washington University – St. Louis, US
- Jen Rogers
Tufts University – Medford, US
- Sheeba Samuel
Friedrich-Schiller-Universität Jena, DE
- Marc Streit
Johannes Kepler Universität Linz, AT
- Tanja Tornede
Leibniz Universität Hannover, DE
- Cagatay Turkay
University of Warwick – Coventry, GB
- Emily Wall
Emory University – Atlanta, US
- Kai Xu
University of Nottingham, GB